Contributors

0-8493-0052-5/00/$0.00+$.50

References

J. Anderson and M. Stonebraker. Sequoia 2000 Metadata Schema for Satellite Images, in [Klaus and Sheth, 1994]. 1994.

T. Berners-Lee et al. World-Wide Web: The Information Universe. *Electronic Networking: Research, Applications and Policy*, 1(2), 1992.

T. Berners-Lee, James Hendler, and Ora Lassila. The Semantic Web. *Scientific American*, May, 2001.

K. Bohm and T. Rakow. Metadata for Multimedia Documents, in [Klaus and Sheth, 1994]. 1994.

S. Boll, W. Klas, and A. Sheth. Overview on Using Metadata to manage Multimedia Data. In A. Sheth and W. Klas, editors, *Multimedia Data Management*. McGraw-Hill, 1998.

T. Bray, J. Paoli, and C. M. Sperberg-McQueen. Extensible markup language (xml) 1.0. http://www.w3.org/TR/REC-xml.

F. Chen, M. Hearst, J. Kupiec, J. Pederson, and L. Wilcox. Metadata for Mixed-Media Access, in [Klaus and Sheth, 1994]. 1994.

C. Collet, M. Huhns, and W. Shen. Resource Integration using a Large Knowledge Base in Carnot. *IEEE Computer*, December 1991.

CPT. Current procedural terminology. http://www.ama-assn.org/ama/pub/category/3113.html.

DAML+OIL. The DARPA Agent Markup Language. http://www.daml.org/.

S. Deerwester, S. Dumais, G. Furnas, T. Landauer, and R. Hashman. Indexing by Latent Semantic Indexing. *Journal of the American Society for Information Science*, 41(6), 1990.

B. Falkenhainer et al. CML: A Compositional Modeling Language, 1994. DRAFT.

M. Forster and P. Mevert. A tool for network modeling. *European Journal of Operational Research*, 72, 1994.

R. Fourer, D. M. Gay, and B. W. Kernighan. AMPL: A Mathematical Programming Language. Technical Report 87-03, Department of Industrial Engineering and Management Sciences, Northwestern University, 1987.

U. Glavitsch, P. Schauble, and M. Wechsler. Metadata for Integrating Speech Documents in a Text Retrieval System, in [Klaus and Sheth, 1994]. 1994.

T. Gruber. A translation approach to portable ontology specifications. *Knowledge Acquisition, An International Journal of Knowledge Acquisition for Knowledge-Based Systems*, 5(2), June 1993.

M. Gyssens et al. A Graph-Oriented Object Database Model. *IEEE Transactions on Knowledge and Data Engineering*, 6(4), 1994.

HL7. The health level seven standard. http://www.hl7.org.

ICD. The international classification of diseases, 9th revision, clinicial modification.

0-8493-0052-5/00/$0.00+$.50

http://www.cdc.gov/nchs/about/otheract/icd9/abticd9.htm.

R. Jain. Semantics in Multimedia Systems. *IEEE Multimedia*, 1(2), 1994.

R. Jain and A. Hampapur. Representations of Video Databases, in [Klaus and Sheth, 1994]. 1994.

B. Kahle and A. Medlar. An Information System for Corporate Users: Wide Area Information Servers. *Connexions - The Interoperability Report*, 5(11), November 1991.

V. Kashyap and A. Sheth. Semantics-based Information Brokering. In *Proceedings of the Third International Conference on Information and Knowledge Management (CIKM)*, November 1994.

V. Kashyap and A. Sheth. Semantic Heterogeneity: Role of Metadata, Context and Ontologies. In M. Papazoglou and G. Schlageter, editors, *Cooperative Information Systems: Current Trends and Directions*. 1997.

D. Kendrick and A. Meeraus. GAMS: An introduction. Technical report, Development Research Department, The World Bank, 1987.

Y. Kiyoki, T. Kitagawa, and T. Hayama. A meta-database System for Semantic Image Search by a Mathematical Model of Meaning, in [Klaus and Sheth, 1994]. 1994.

W. Klaus and A. Sheth. Metadata for digital media. *SIGMOD Record, special issue on Metadata for Digital Media*, W. Klaus, A. Sheth, eds., 23(4), December 1994.

O. Lassila and R. R. Swick. Resource description framework (rdf) model and syntax specification. http://www.w3.org/TR/REC-rdf-syntax/.

D. Lindbergh, B. Humphreys, and A. McCray. The Unified Medical Language System. *Methods Inf. Med.*, 32(4), 1993. http://umlsks.nlm.nih.gov.

LOINC. The logical observation identifiers names and codes database. http://www.loinc.org.

A. McCray and S. Nelson. The representation of meaning in the UMLS. *Methods Ind. Med.*, 34(1-2):193–201, 1995.

A. McCray, S. Srinivasan, and A. Browne. Lexical methods for managing variation in biomedical terminologies. In *Proceedings of the Annual Symposium on Computers in Applied Medical Care*, 1994.

MEDLINE. The PubMed MEDLINE system. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=PubMed.

E. Mena, V. Kashyap, A. Sheth, and A. Illarramendi. OBSERVER: An approach for query processing in global information systems based on interoperation across pre-existing ontologies. In *Proceedings of the First IFCIS International Conference on Cooperative Information Systems (CoopIS '96)*, June 1996.

S. J. Nelson, W. D. Johnston, and B. L. Humphreys. Relationships in Medical Subject Headings (MeSH). In C. A. Bean and R. Green, editors, *Relationships in the Organization of Knowledge*. Kluwer Academic Publishers, 2001.

L. Neustadter. A Formalization of Expression Semantics for an Executable Modeling Language. In *Proceedings of the Twenty Seventh Annual Hawaii International Conference on System Sciences*, 1994.

V. Ogle and M. Stonebraker. Chabot: Retrieval from a Relational database of images. *IEEE Computer, special issue on Content-Based Image Retrieval Systems*, 28(9), 1995.

J. Ordille and B. Miller. Distributed Active Catalogs and Meta-Data Caching in Descriptive Name Services. In *Proceedings of the 13th International Conference on Distributed Computing Systems*,

May 1993.

OWL. The Web Ontology Language. http://www.w3.org/TR/owl_guide/.

G. Salton. *Automatic text processing*. Addison-Wesley, 1989.

E. Sciore, M. Siegel, and A. Rosenthal. Context Interchange using Meta-Attributes. In *Proceedings of the CIKM*, 1992.

SGML. The Standard Generalized Markup Language. http://www.w3.org/MarkUp/SGML/.

K. Shah and A. Sheth. Logical Information Modeling of Web-accessible Heterogeneous Digital Assets. In *Proceedings of the IEEE Advances in Digital Libraries (ADL) Conference*, April 1998.

A. Sheth and V. Kashyap. Media-independent Correlation of Information. What? How? In *Proceedings of the First IEEE Metadata Conference*, April 1996. http://www.computer.org/conferences/meta96/sheth/index.html.

A. Sheth, V. Kashyap, and W. LeBlanc. Attribute-based access of Heterogeneous Digital Data. In *Proceedings of the Workshop on Web Access to Legacy Data, Fourth International WWW Conference*, December 1995.

L. Shklar, K. Shah, and C. Basu. The InfoHarness Repository Definition Language. In *Proceedings of the Third International WWW Conference*, May 1995a.

L. Shklar, K. Shah, C. Basu, and V. Kashyap. Modelling Heterogeneous Information. In *Proceedings of the Second International Workshop on Next Generation Information Technologies (NGITS '95)*, June 1995b.

L. Shklar, A. Sheth, V. Kashyap, and K. Shah. Infoharness: Use of Automatically Generated Metadata for Search and Retrieval of Heterogeneous Information. In *Proceedings of CAiSE '95*, June 1995c. Lecture Notes in Computer Science, #932.

L. Shklar, S. Thatte, H. Marcus, and A. Sheth. The InfoHarness Information Integration Platform. In *Proceedings of the Second International WWW Conference*, October 1994.

K. Shoens, A. Luniewski, P. Schwartz, J. Stamos, and J. Thomas. The Rufus System: Information Organization for Semi-Structured Data. In *Proceedings of the 19th VLDB Conference*, September 1993.

Snomed. The systematized nomenclature of medicine. http://www.snomed.org.

SVIP. The semantic vocabulary interoperation project. http://cgsb2.nlm.nih.gov/ kashyap/projects/SVIP/.

J. H. Taylor. Towards a Modeling Language Standard for Hybrid Dynamical Systems. In *Proceedings of the 32nd Conference on Decision and Control*, 1993.

1

Information Modeling on the Web: The Role of Metadata, Semantics and Ontologies

Vipul Kashyap
Lister Hill National Center for Biomedical Communications, National Library of Medicine, NIH

CONTENTS

The web consists of huge amounts of data available in a variety of digital forms stored in thousands of repositories. Approaches that use the semantics of information captured in the metadata extracted from the data are being viewed as an appealing approach, especially in the context of the *Semantic Web* effort.

We present in this chapter a discussion on approaches adopted for metadata-based information modeling on the web. Various types of metadata developed by researchers for different media are reviewed and classified with respect to the extent they model data or information content. The reference terms and the ontology of the metadata are classified *wrt* their dependence on the application domain. We discuss approaches of using the metadata to represent the context of the information request, the interrelationships between the various pieces of the data and exploit them for search, browsing and querying the information. Issues related to the use of terminologies and ontologies such as establishing and maintaining terminological commitments, and their role in metadata design and extraction are also discussed. Modeling languages and formats, including the most recent ones, such as the Resource Description Format (RDF) and the DARPA Agent Markup Language (DAML+OIL) are also discussed in this context.

1.1 Introduction

The World Wide Web [Berners-Lee et al., 1992] consists of huge amounts of digital data in a variety of structured, unstructured (e.g., image) and sequential (e.g., audio, video) formats that are either stored as web data directly manipulated by web servers, or retrieved from underlying database and content management systems and served as dynamically generated web content. Whereas content management systems support creation, storage and access functions in the context of the content managed by them, there is a need to support correlation across different types of digital formats in *media independent, content-based* manner.

Information relevant to a user or application need may be stored in multiple forms (e.g., structured data, text, image, audio and video) in different repositories and web sites. Responding to a

0-8493-0052-5/00/$0.00+$.50

user's information request typically requires correlation of such information across multiple forms and representations. There is a need for association of various pieces of data either by pre-analysis by software programs or dynamic correlation of information in response to an information request. Common to both the approaches is the ability to describe the *semantics* of the information represented by the underlying data.

The use of semantic information to support correlation of heterogeneous representations of information is one of the aims of the current *Semantic Web* effort [Berners-Lee et al., 2001]. This capability of modeling information at a semantic level both across different types of structured data (e.g., in data warehouses) and across different types of multimedia content, is missing on the current web and has been referred to as the "semantic bottleneck" [Jain, 1994]. Machine understandable metadata and standardized representations thereof form the foundation of the Semantic Web. The *Resource Description Framework (RDF)* [Lassila and Swick] and *XML* [Bray et al.] based specifications are currently being developed in an effort to standardize the formats for representing metadata. It is proposed that the vocabulary terms used to create the metadata will be chosen from third party ontologies available from the web. Standardized specifications for representing ontologies include XML and RDF schemas, *DARPA Agent Markup Language* (DAML+OIL) [DAML+OIL] and the *Web Ontology Language (OWL)* [OWL].

In this chapter, we present issues related to the use of *metadata, semantics and ontologies* for modeling information on the web organized in a three level framework (Figure 1.1):

- The middle level represents the **metadata** component involving the use of metadata descriptions to capture the information content of data stored in websites and repositories. Intensional descriptions constructed from metadata, are used to abstract from the structure and organization of data and specify relationships across pieces of interest.
- The top level represents the **ontology** component, involving terms (concepts, roles) in domain specific ontologies used to characterize metadata descriptions. These terms capture pieces of domain knowledge that describe relationships between data items (via association with the terms) across multiple repositories, enabling semantic interoperability.

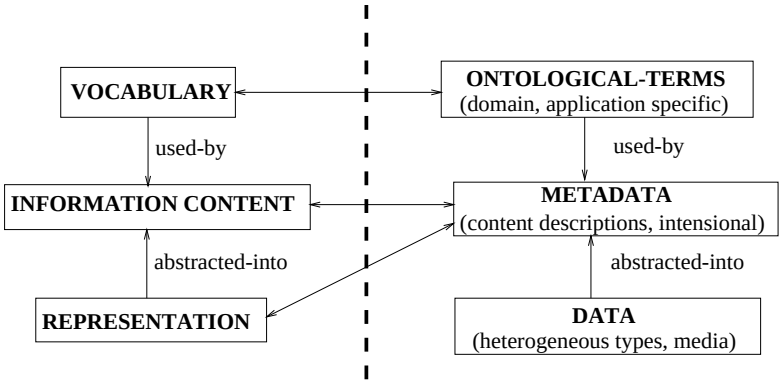


FIGURE 1.1
Key issues for information modeling

The organization of this chapter is as follows. In Section 1.2, we discuss a definition of metadata, with various examples. A classification of metadata based on the information content they capture is presented along with its role in modeling information. In Section 1.3, we discuss how metadata

expressions can be used to model interrelationships between various pieces of information within a dataset and across multiple datasets. We also present an account of various modeling and markup languages that may be used to model the information represented in the data. Finally, in Section 1.4, we present issues related to the use of reference terms and ontological concepts for creating metadata descriptions. Section 1.5 presents the conclusions.

1.2 What is Metadata?

Metadata in its most general sense is defined as data or information about data. For structured databases, the most common example of metadata is the schema of the database. However with the proliferation of various types of multimedia data on the web, we shall refer to an expanded notion of metadata of which the schema of structured databases is a (small) part. Metadata may be used to store derived properties of media useful in information access or retrieval. They may describe or be a summary of the information content of the data described in an intensional manner. They may also be used to represent properties of or relationships between individual objects of heterogeneous types and media. Figure 1.1 illustrates the components for modeling information on the web. Metadata is the pivotal idea on which both the (ontology and metadata) components depend. The function of the metadata descriptions is two-fold:

- To enable the abstraction of representational details such as the format and organization of data, and capture the information content of the underlying data independent of representational details. These expressions may be used to represent useful relationships between various pieces of data within a repository or web site.
- To enable representation of domain knowledge describing the information domain to which the underlying data belongs. This knowledge may then be used to make inferences about the underlying data to determine the *relevance* and indentify relationships across data stored in different repositories and web sites.

We now discuss issues related to metadata from two different perspectives identified in [Boll et al., 1998], viz., the usage of metadata in various applications, and the information content captured by the metadata.

1.2.1 Metadata usage in various applications

We discuss a set of application scenarios that require functionality for manipulation and retrieval of digital content that are relevant to the web. The role of metadata especially in the context of modeling information to support this functionality is discussed.

Navigation, Browsing and Retrieval from Image Collections An increasing number of applications, such as those in healthcare, maintain large collections of images. There is a need for semantic content based navigation, browsing, and retrieval of images. An important issue is to associate a user's semantic impression with the images, e.g., image of a brain tumor. This requires knowledge of spatial content of the image, and the way it changes or evolves over time, which can be represented as metadata annotations.

Video In many applications relevant to news agencies, there exist collections of video footage which need to be searched based on semantic content, e.g., videos containing field goals in a soccer game. This gives rise to the same set of issues as described above, such as the change

in the spatial positions of various objects in the video images (spatial evolution). However, there is a temporal aspect to videos that was not captured above. Sophisticated time-stamp based schemes can be represented as a part of the metadata annotations.

Audio and Speech Radio stations collect many, if not all of their important and informative programs, such as radio news, in archives. Parts of such programs are often reused in other radio broadcasts. However, to efficiently retrieve parts of radio programs, it is necessary to have the right metadata generated from, and associated with, the audio recordings. An important issue here is capturing in text, the essence of the audio, in which vocabulary plays a central role. Domain specific vocabularies can drive the metadata extraction process making it more efficient.

Structured Document Management As the publishing paradigm is shifting from popular desktop publishing to database-driven web-based publishing, processing of structured documents becomes more and more important. Particular document information models, such as SGML [SGML] and XML, introduce structure and content-based metadata. Efficient retrieval is achieved by exploiting document structure, as the metadata can be used for indexing, which is essential for quick response times. Thus, queries asking for documents with a title containing "Computer Science" can be easily optimized.

Geographic and Environmental Information Systems These systems have a wide variety of users that have very specific information needs. Information integration is a key requirement, which is supported by provision of descriptive information to end users and information systems. This involves issues of capturing descriptions as metadataand reconciling the different vocabularies used by the different information systems in interpreting the descriptions.

Digital Libraries Digital libraries offer a wide range of services and collections of digital documents, and constitute a challenging application area for the development and implementation of metadata frameworks. These frameworks are geared towards description of collections of digital materials such as text documents, spatially referenced data sets, audio, and video. Some frameworks follow the traditional library paradigm with metadata like subject headings [Nelson et al., 2001] and thesauri [Lindbergh et al., 1993].

Mixed Media Access This is an approach which allows queries to be specified independent of the underlying media types. Data corresponding to the query may be retrieved from different media such as text and images, and "fused" appropriately before being presented to the user. Symbolic metadata descriptions may be used to describe information from different media types in a uniform manner.

1.2.2 Metadata: A means for modeling information

We now characterize various types of metadata based on the amount of information content they capture, and present a classification of various types of metadata used by researchers (Table 1.2.2).

Content Independent Metadata This type of metadata captures information that does not depend on the content of the document with which it is associated. Examples of this type of metadata are location, modification-date of a document and type-of-sensor used to record a photographic image. There is no information content captured by these metadata but these might still be useful for retrieval of documents from their actual physical locations, and for checking whether the information is current or not. This type of metadata helps to encapsulate information into units of interest, and organizes their representation within an object model.

Content Dependent Metadata This type of metadata depends on the content of the document it is associated with. Examples of content dependent metadata are size of a document, max-colors, number-of-rows, and number-of-columns of an image. These type of metadata typically capture representational and structural information, and provide support for browsing and navigation of the underlying data. Content dependent metadata can be further sub-divided as follows:

Direct Content-based Metadata This type of metadata is based directly on the contents of a document. A popular example of this is full-text indices based on the document text. Inverted tree and document vectors are examples of this type of metadata. *Media specific metadata* such as color, shape, and texture are typically direct content-based metadata.

Content-descriptive Metadata This type of metadata describes information in a document without directly utilizing its contents. An example of this metadata is textual annotations describing the contents of an image. This metadata comes in two flavors:

Domain Independent Metadata These metadata capture information present in the document independent of the application or subject domain of the information, and are primarily structural in nature. They often form the basis of indexing the document collection to enable faster retrieval. Examples of these are C/C++ parse trees and HTML/SGML document type definitions. Indexing a document collection based on domain independent metadata may be used to improve retrieval efficiency.

Domain Specific Metadata Metadata of this type is described in a manner specific to the application or subject domain of the information. Issues of vocabulary become very important in this case, as the metadata terms have to be chosen in a domain specific manner. This type of metadata, which helps abstract out representational details and capture information meaningful to a particular application or subject domain, is Domain Specific Metadata. Examples of such metadata are relief, land-cover from the geographical information domain and medical subject headings (MeSH) from the medical domain. In the case of structured data, the database schema is an example of such metadata. These type of metadata can be further categorized as:

Intra-domain Specific Metadata These type of metadata capture relationships and associations between data within the context of the same information domain. For example, the relationship between the CEO and his corporation is captured within a common information domain, viz., the business domain.

Inter-domain Specific Metadata These type of metadata capture relationships and associations between data across information domains. For example, the relationship between (medical) instrument and (legal) instrument spans across the medical and legal information domains.

Vocabulary for Information Content Characterization Domain Specific Metadata can be constructed from terms in a controlled vocabulary of terms and concepts, e.g. the biomedical vocabularies available in the Unified Medical Language System (UMLS) [Lindbergh et al., 1993], or a domain specific ontology, describing information in an application or subject domain. Thus, we view ontologies as metadata, which themselves can be viewed as a vocabulary of terms for construction of more domain specific metadata descriptions.

Crisp vs Fuzzy Metadata This is an orthogonal dimension for categorization. Some of the metadata referred to above are fuzzy in nature and are modeled using statistical methods, e.g., document vectors. On the other hand other metadata annotations might be of a crisp nature, e.g., author name.

Metadata	Media/Metadata Type
Q-Features	Image, Video/Domain Specific
R-Features	Image, Video/Domain Independent
Impression Vector	Image/Content Descriptive
NDVI, Spatial Registration	Image/Domain Specific
Speech feature index	Audio/Direct Content-based
Topic change indices	Audio/Direct Content-based
Document Vectors	Text/Direct Content-based
Inverted Indices	Text/Direct Content-based
Content Classification Metadata	MultiMedia/Domain Specific
Document Composition Metadata	MultiMedia/Domain Independent
Metadata Templates	Media Independent/Domain Specific
Land-Cover, Relief	Media Independent/Domain Specific
Parent-Child Relationships	Text/Domain Independent
Contexts	Structured Databases/Domain Specific
Concepts from Cyc	Structured Databases/Domain Specific
User's Data Attributes	Text, Structured Databases/Domain Specific
Medical Subject Headings	Text Databases/Domain Specific
Domain Specific Ontologies	Media-Independent/Domain Specific

Metadata for Digital Media

In the above table we have surveyed different types of metadata used by various researchers. Q-Features and R-Features were used for modeling image and video data [Jain and Hampapur, 1994]. Impression vectors were generated from text descriptions of images [Kiyoki et al., 1994]. NDVI and spatial registration metadata were used to model geo-spatial maps, primarily of different types of vegetation [Anderson and Stonebraker, 1994]. Interesting examples of mixed media access are the speech feature index [Glavitsch et al., 1994] and topic change indices [Chen et al., 1994]. Metadata capturing information about documents are document vectors [Deerwester et al., 1990], inverted indices [Kahle and Medlar, 1991], document classification and composition metadata [Bohm and Rakow, 1994] and parent-child relationships (based on document structure) [Shklar et al., 1995c]. Metadata Templates [Ordille and Miller, 1993] have been used for information resource discovery. Semantic metadata such as contexts [Sciore et al., 1992; Kashyap and Sheth, 1994], land-cover, relief [Sheth and Kashyap, 1996], Cyc concepts [Collet et al., 1991], concepts from domain ontologies [Mena et al., 1996] have been constructed from well defined and standardized vocabularies and ontologies. Medical Subject headings [Nelson et al., 2001] are used to annotate biomedical research articles in MEDLINE [MEDLINE]. These are constructed from biomedical vocabularies available in the UMLS [Lindbergh et al., 1993]. An attempt at modeling user attributes is presented in [Shoens et al., 1993]. The above discussion suggests that domain specific metadata capture information which is more meaningful with respect to a specific application or a domain. The information captured by other types of metadata primarily reflect the format and organization of underlying data.

1.3 Metadata Expressions: Modeling Information Content

We presented in the previous section, different types of metadata that capture information content to different extents. Metadata has been used by a wide variety of researchers in various contexts for different functionality relating to retrieval and manipulation of digital content. We now discuss approaches for combining metadata to create information models based on the underlying data. There are two broad approaches:

- Use of content and domain independent metadata to encapsulate digital content within an infrastructural object model.
- Use of domain specific metadata to specify existing relationships within the same content collection or across collections.

We present both these approaches, followed by a brief survey of modeling and markup languages that have been used

1.3.1 The InfoHarness System: Metadata-based Object Model for Digital Content

We now discuss the InfoHarness [Shklar et al., 1994, 1995c,a,b; Sheth et al., 1995] system, which has been the basis of many successful research projects and commercial products. The main goal of InfoHarness is to provide a uniform access to information independent of the formats, location and organization of the information in the individual information sources. We discuss how content-independent metadata (e.g., type, location, access rights, owner, creation date, etc.) may be used to encapsulate the underlying data and media heterogeneity and represent information in an object model. We then discuss how the information spaces might be logically structured and discuss an approach for an interpreted modeling language.

1.3.1.1 Metadata for encapsulation of information

Representational details are abstracted out of the underlying data and metadata is used to capture information content. This is achieved by encapsulation of the underlying data into units of interest called information units, and extraction of metadata describing information of interest. The object representation is illustrated in Figure 1.2 and is discussed below.

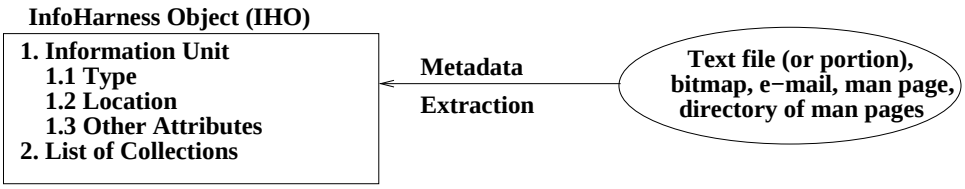


FIGURE 1.2
Metadata Encapsulation in InfoHarness

A metadata entity that is associated with the lowest level of granularity of information available to InfoHarness is called an *information unit* (IU). An IU may be associated with a file (e.g., a Unix man page or help file, a Usenet news item), a portion of a file (e.g., a C function or a database table), a set of files (e.g., a collection of related bitmaps), or any request for the retrieval of data from an external source (e.g., a database query). An InfoHarness Object (IHO) may be one of the following:

1. A single information unit.
2. A collection of InfoHarness objects (either indexed, or non-indexed).
3. A single information unit and a non-indexed collection of InfoHarness objects.

Each IHO has a unique object identifier that is recognized and maintained by the system. An IHO that encapsulates an IU contains information about the location of data, retrieval method, and any parameters needed by the method to extract the relevant portion of information. For example, an

IHO associated with a C function will contain the path information for the .c file that contains the function, the name and location of the program that knows how to extract a function from a .c file, and the name of the function to be passed to this program as a parameter. In addition each IHO may contain an arbitrary number of attribute-value pairs for attribute-based access to the information. An InfoHarness Repository (IHR) is a collection of IHOs. Each IHO (known as the *parent*) that encapsulates a collection of IHOs stores unique object identifiers of the members of the collection. We refer to these members as *children* of the IHO. IHOs that encapsulate indexed collections store information about the location of both the index and the query method.

1.3.1.2 Logical Structuring of the Information Space

We now discuss the various types of logical structure that can be imposed on the content in the context of the functionality enabled by such a structuring. This structuring is enabled by the extraction of the different kinds of metadata discussed above.

Consider the scenario illustrated in Figure 1.3. Case I depicts the actual physical distribution of the various types of documents required in a large software design project. The different documents are spread all over the file system as a result of different members of the project putting the files where they deemed appropriate. Appropriate metadata extractors pre-process these documents and store important information like *type* and *location* and establish appropriate parent-child relationships. Case II illustrates the desired logical view seen by the user. Information can be browsed according to units of interest as opposed to browsing the information according to the physical organization in the underlying data repositories.

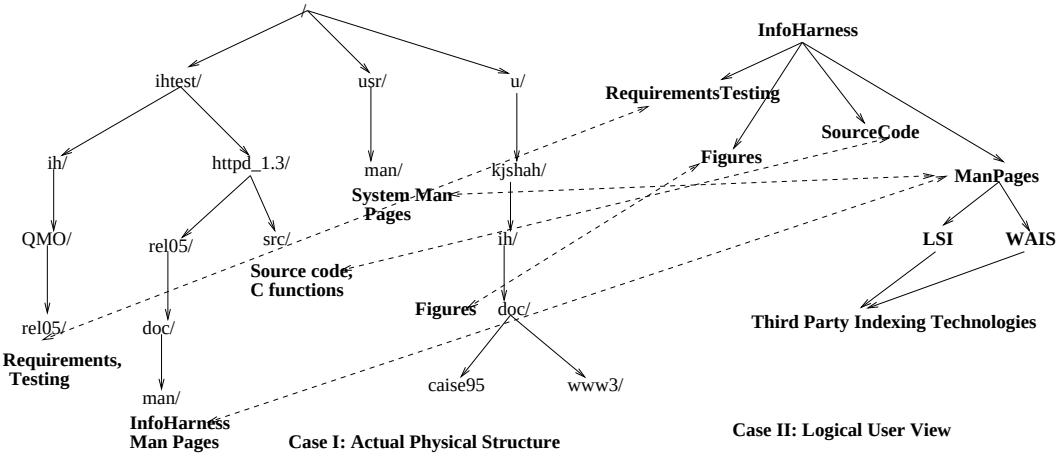


FIGURE 1.3
Logical structuring of the Information Space

One of the key capabilities enabled by the logical structuring is the ability to seamlessly plug in third-party indexing technologies to index document collections. This is illustrated in Figure 1.3, Case II where the same set of documents are indexed using different third party indexing technologies. Each of these document collections so indexed can be now queried using a *keyword-based query* without the user having to worry about the details of the underlying indexing technology.

Attribute-based access provides a powerful complementary or alternative search mechanism to traditional content-based search and access [Sheth et al., 1995]. While attribute-based access can provide better *precision* [Salton, 1989], it can be more complex as it requires that appropriate at-

tributes have been identified and the corresponding metadata instantiated before accessing data.

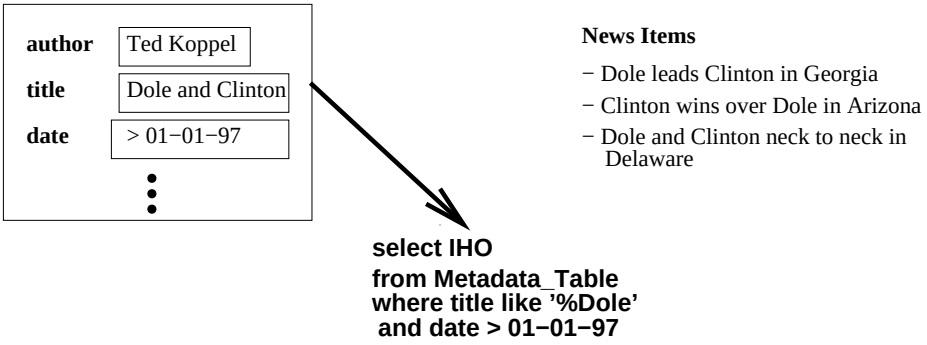


FIGURE 1.4
Attribute Based Access in InfoHarness

In Figure 1.4 we illustrate an example of attribute-based access in InfoHarness. Attribute-based queries by the user result in SQL queries to the metadata repository and result in retrieval of the news items which satisfy the conditions specified. The advantages of attribute-based access are:

Enhance the semantics of the keywords When a user presents a keyword (e.g., "Ted Koppel") as the value of an attribute (e.g., author) there are more constraints on the keyword as compared to when it appears by itself in a keyword-based query. This improves the precision of the answer.

Attributes can have associated types The attribute submission date could have values of type date. Simple comparison operators ($<$, $>$, $\leq$, $\geq$) can now be used for specifying constraints.

Querying content independent information One cannot query content independent information like modification date using keyword based access as such information will never be available from the analysis of the content of the document.

1.3.1.3 IRDL: A Modeling Language for generating the Object Model

The creation of an IHR amounts to the generation of metadata objects that represent IHOs and to the indexing of physical information encapsulated by members of indexed collections. The IHR can either be generated manually by writing metadata extractors or created automatically by interpreting IRDL (InfoHarness Repository Definition Language) statements. A detailed discussion of IRDL can be found in [Shklar et al., 1995a] and its use in modeling heterogeneous information is discussed in [Shklar et al., 1995b]. There are three main IRDL commands:

Encapsulate This command takes as input information about the *type* and *location* of physical data and returns a set of IHOs, each of which encapsulates a piece of data. Boundaries of these pieces are determined by the type. For example, in the case of e-mail, a set IHOs, each of which is associated with a separate mail message is returned.

Group This command generates an IHO associated with a collection and establishes parent-child relationships between the collection IHO and the member IHOs. In case, a perimeter indicating the indexing technology is specified, an index on the physical data associated with the member IHOs is created.

Merge This command takes as input an IHO and associated references and creates a composite IHO.

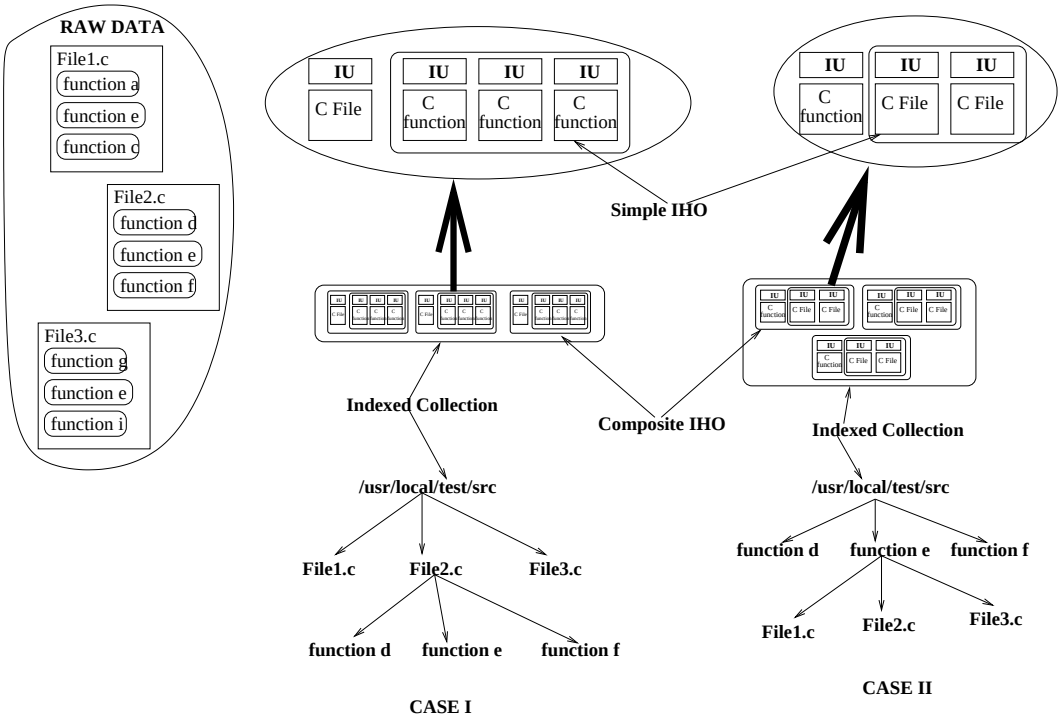


FIGURE 1.5
Object Model Generation for a C program

We explain the model generation process by discussing an example for C programs (Figure 1.5). The steps that generate the model displayed in Figure 1.5, Case I are as follows:

1. For each C file do the following:
   (a) Create simple IHOs that encapsulate individual functions that occur in this file.
   (b) Create a composite IHO that encapsulates the file and points to IHOs created in step 1.1.
2. Create an indexed collection of the composite IHOs created in step 1, using LSI for indexing physical data.

The IRDL statements that generate the model discussed above are as below.

```
BEGIN
     COLLTYPE LSI;
     DATATYPE TXT, C;
     VAR IHO: File_IHO, LSI_Collection;
     VAR SET IHO: File_IHO_SET, Function_IHO_SET;
     File_IHO_SET = ENCAP TXT "/usr/local/test/src";
     FORALL File_IHO IN File_IHO_SET
     {
```

```
            Function_IHO_SET = ENCAP C File_IHO;
            File_IHO = COMBINE IHO Function_IHO_SET;
            WRITE File_IHO, Function_IHO_SET;
    }
    LSI_Collection = INDEX LSI File_IHO_SET "/usr/local/db/c";
    WRITE LSI_Collection;
END
```

This is another example of logical structuring using parent-child relationships to set up different logical views of the same underlying physical space. Case I (Figure 1.5) illustrates the case where a directory containing C code is viewed as a collection of C files each of which is a collection of C functions. Case II (Figure 1.5) on the other hand illustrates the case where the directory is viewed as a collection of C functions each of which is a collection of the C files in which it appears.

1.3.2 Metadata-based Logical Semantic Webs

The Web as it exists today is a graph of information artifacts and resources, where graph nodes are represented by embedded HREF tags. These tags enable linking of related (or unrelated) web artifacts. This web is very suitable for browsing but provides little or no direct help for searching, information gathering or analysis. Web crawlers and search engines try to impose some sort of an order by building indices on top of web artifacts, which are primarily textual. For example, a keyword query may be viewed as imposing a correlation (logical relationship) at a very basic (limited) level between the artifacts that make up the result set for that query. For example, let's say a search query "Bill Clinton" (Q) retrieves http://www.billclinton.com (Resource1) and http://www.whitehouse.gov/billclinton.html (Resource2). An interesting viewpoint would be that Resource1 and Resource2 are "correlated" with each other through the query Q. The above may be represented graphically with Resource1 and Resource2 represented as nodes linked by an edge labeled by the string corresponding to Q. Metadata is the key to this correlation. The keyword index (used to process keyword queries) may be conceptually viewed as content-dependent metadata, and the keywords in the query as specific resource descriptors for the index, the evaluation of which would result in a set of linked or correlated resources.

We discussed in the previous section, the role played by metadata in encapsulating digital content into an object model. An approach using an interpreted modeling language for metadata extraction and generation of the object model was presented. We now present a discussion on how a metadata based formalism, the metadata reference link (MREF) [Sheth and Kashyap, 1996; Shah and Sheth, 1998] can be used to enable semantic linking correlation, an important pre-requisite for building logical semantic webs. MREF is a generalization of the <A HREF> construct used by the current web to specify links and is defined as follows.

- <A MREF KEYWORDS=[keyword-list] THRESHOLD=[real] >Document Description< /A>
- <A MREF ATTRIBUTES([attr-value-pair-list])>Document Description< /A>

Different types of correlation are enabled based on the type of metadata that is used. We now present examples of correlation with the help of examples.

1.3.2.1 Content Independent Correlation

This type of correlation arises when content independent metadata (e.g., the location expressed as a URL) is used to establish the correlation. The correlation is typically media independent as content independent metadata typically does not depend on media characteristics. In this case, the correlation is done by the designer of the document as illustrated in the following example:

```
<TITLE>A Scenic Sunset at Lake Tahoe</TITLE>
Lake Tahoe is a very popular tourist spot and
<A HREF="http://www1.server.edu/lake-tahoe.txt">some interesting
facts</A> are available here. The scenic beauty of Lake Tahoe can
be viewed in this photograph:
<center>
<IMG ALIGN=MIDDLE SRC="http://www2.server.edu/lake-tahoe.img">
</center>
```

The correlation is achieved by using physical links and without using any higher level specification mechanism. This is predominantly the type of correlation found in the HTML documents on the World Wide Web [Berners-Lee et al., 1992].

1.3.2.2 Correlation using Direct Content-based Metadata

We present below an example based on a query in [Ogle and Stonebraker, 1995] to demonstrate a correlation involving attribute based metadata. One of the attributes is color which is a **media specific** attribute. Hence we view this interesting case of correlation as **media specific** correlation.

```
<TITLE>Scenic Waterfalls</TITLE>
Some interesting
<A MREF ATTRIBUTES(keyword="scenic waterfalls"; color="blue")>
information on scenic waterfalls</A> is available here.
```

1.3.2.3 Correlation using Content-descriptive Metadata

In [Kiyoki et al., 1994], keywords are associated with images and a full-text index is created on the key word descriptions. Since the keywords describe the contents of an image, we consider these as **content-descriptive** metadata. Correlation can now be achieved by querying the collection of image documents and text documents using the same set of keywords as illustrated in this example:

```
<TITLE>Scenic Natural Sights</TITLE>
Some interesting
<A MREF KEYWORDS="scenic waterfall mountain",THRESH=0.9>
information on Lake Tahoe</A> is available here.
```

This type of correlation is more meaningful than content independent correlation. Also the user has more control over the correlation, as he may be allowed to change the thresholds and the keywords. The keywords used to describe the image are media independent and hence correlation is achieved in a media independent manner.

1.3.2.4 Domain Specific Correlation

To better handle the information overload on the fast growing global information infrastructure, there needs to be support for correlation of information at a higher level of abstraction independent of the medium of representation of the information [Jain, 1994]. Domain specific metadata, which is necessarily media independent needs to modeled. Let us consider the domain of a Site Location and Planning application supported by a Geographic Information System and a correlation query illustrated in the following example:

```
<TITLE>Site Location and Planning</TITLE>
To identify potential locations for a future shopping mall, we present
below all regions having a population greater than 500 and area greater
than 50 sq feet having an urban land cover and moderate relief
<A MREF ATTRIBUTES(population &gt; 500; area &gt; 50; regiontype = block;
landcover = urban; relief = moderate)> can be viewed here</A>
```

The processing of the above query results in the structured information (area, population) and the map of the regions satisfying the above constraints being included in the HTML document. The query processing system will have to map these attributes to image processing and other SQL-based routines to retrieve and present the results.

1.3.2.5 Example: RDF representation of MREF

These notions of metadata-based modeling are fundamental to the notion of the emerging semantic web [Berners-Lee et al., 2001]. Semantic web researchers have focused on markup languages for representing machine understandable metadata. We now present a representation of the example listed above using the RDF markup language.

```
<HEAD>
<OBJECT declare id="mall-loc" type="application/x-mref"
      data="<?namespace href="http://www.foo.com/SitePlanning" as="SP"?>
      <?namespace href="http://www.w3.org/schemas/rdf-schema" as="RDF"?>
            <RDF:serialization>
            <RDF:bag id="MREF:mall-loc>
<SP:attribute>
      <RDF:resource id="constraint_001">
    <SP:name>population</SP:color>
    <SP:type>number</SP:type>
                    <SP:operator>greater</SP:operator>
    <RDF:PropValue>500</RDF:PropValue>
      </RDF:resource>
      </SP:attribute>
      ....
        </RDF:bag>
          </RDF:serialization>">
</OBJECT>
</HEAD>
<BODY>
To identify potential locations for a future shopping mall, all regions
having a population greater than 500 and area greater than 50 acres
having an urban land cover and moderate relief
<OBJECT classid="http://www.foo.com/sp.mref"
standby="Loading MREF..." data="#mall-loc">can be viewed here.</OBJECT>
</BODY>
```

1.3.3 Modeling Languages and Markup Standards

The concept of a simple, declarative language to support modeling is not new. Although modeling languages borrow from the classical hierarchical, relational and network approaches, a number of them incorporate and extend the relational model. The languages examined below may be categorized as:

Algebraic model formulation generators AMPL [Fourer et al., 1987], GAMS [Kendrick and Meeraus, 1987] and GEML [Neustadter, 1994] belong to this group.

Graphical model generators GOOD [Gyssens et al., 1994] and GYNGEN [Forster and Mevert, 1994] belong to this group.

Hybrid/compositional model generators These languages have an underlying representation based on mathematical and symbolic properties, e.g., CML [Falkenhainer et al., 1994] and SHSML [Taylor, 1993].

GOOD attempts to provide ease of high-level conceptualizing and manipulation of data. Sharing similarities with GOOD, GYNGEN focuses on process modeling by capturing the semantics underlying planning problems. CML and SHSML facilitate the modeling of dynamic processes. GEML is a language based on sets. and has both primitive and derived data types. Primitive types may be defined by the user or built-in scalars. Derived types are recursive applications of operations such as the cartesian product or sub-typing.

GOOD, GYNGEN, SHSML, and CML all employ graphs for defining structures. For the individual languages, variations arise when determining the role of nodes/edges as representations of the underlying concepts and composing and interconnecting them to produce meaningful representations. GOOD relies on the operations of node addition/deletion, edge addition/deletion, and abstraction to build directed graphs. SHSML and CML are designed specifically to handle data dependencies arising from dynamic processes with time-varying properties.

Structure in CML is domain-theory dependent, defined by a set of top-level forms. Domain theories are composed from components, processes, interaction phenomena, logical relations, etc. The types in this language include symbols, lists, terms composed of lists, sequences, and sets of sequences. The language promotes reuse of existing domain theories to model processes under a variety of conditions.

A host of initiatives have been proposed by the W3C consortium that have a lot in common with the modeling languages listed above. The effort has been to standardize the various features across a wide variety of potential applications on the web and specify markup formats for the same. A list of such markup formats are:

XML XML is a markup language for documents containing structured information. It is a metalanguage for describing markup languages, i.e., it provides a facility to define tags and the structural relationships between them. Since there's no predefined tag set, there can't be any preconceived semantics. All of the semantics of an XML document are defined by specialized instantiations, applications that process XML specifications or by stylesheets. The vocabulary that makes up the tags and associated values may be obtained from ontologies and thesauri possibly available on the web.

XSLT and XPath The Extensible Stylesheet Language Transformations (XSLT) and the XML Path Language (XPath) are essentially languages that support transformation of XML specifications from one language to another.

XML Schema The XML Schema definition language is a markup language that describes and constrains the content of an XML document. It is analogous to the database schema for relational databases and is a generalization of the document type definitions (DTDs).

XQuery The XQuery language is a powerful language for processing and querying XML data. It is analogous to the structured query language (SQL) used in the context of relational databases.

RDF The Resource Description Framework (RDF) is a format for representing machine understandable metadata on the web. It has a graph based data model with *resources* as nodes, *properties* as labeled edges and *values* as nodes.

RDF Schema Though RDF specifies a data model, it doesn't specify the vocabulary, e.g., what properties need to be represented, of the metadata description. These vocabularies (ontologies) are represented using RDF Schema expressions and can be used to constrain the underlying RDF statements.

DAML+OIL The DARPA Agent Markup Language (DAML+OIL) is a more sophisticated specification (compared to RDF Schema) used to capture semantic constraints that might be available in an ontology/vocabulary.

Topic Maps Topic Maps share with RDF the goal of representing relationships amongst data items of interest. A topic map is essentially a collection of topics which are used to describe key concepts in the underlying data repositories (text and relational databases). Relationships to these topics are represented using links also, called *associations*. Links that associate a given topic with the information sources in which it appears are called *occurrences*. Topics are related together independently of what is said about them in the information being indexed. A topic map defines a multidimensional topic space a space in which the locations are topics, and in which the distances between topics are measurable in terms of the number of intervening topics which must be visited in order to get from one topic to another, and the kinds of relationships that define the path from one topic to another, if any, through the intervening topics, if any.

Web Services Web Services are computations available on the web that can be invoked via standardized XML messages. Web Services Description Language (WSDL) describes these services in a repository, the Universal Description, Discovery and Integration Service (UDDI) which can be invoked using the Simple Object Access Protocol (SOAP) specification.

1.4 Ontology: Vocabularies and Reference terms for Metadata

We have discussed in the previous sections, how metadata-based descriptions are an important tool for modeling information on the web. The degree of semantics depends on the nature of these descriptions, i.e., whether they are domain specific. A crucial aspect of creating metadata descriptions is the vocabulary used to create them. The key to utilizing the knowledge of an application domain is identifying the basic vocabulary consisting of terms or concepts of interest to a typical user in the application domain and the interrelationships among the concepts in the ontology.

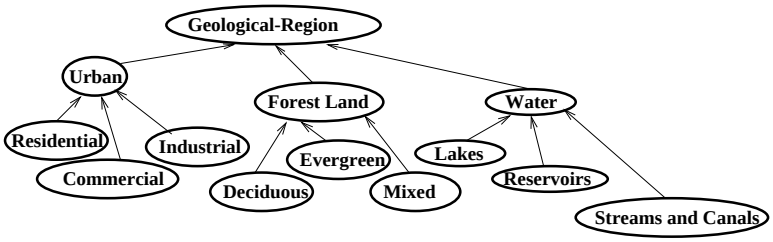
In the course of collecting a vocabulary or constructing an ontology for information represented in a particular media type some concepts or terms may be independent of the application domain. Some of them may be media specific while others might be media independent. There might be some application specific concepts for which interrelationships may be represented. They are typically independent of the media of representation. Information represented using different media types can be modeled with application specific concepts.

1.4.1 Terminological Commitments: Constructing an Ontology

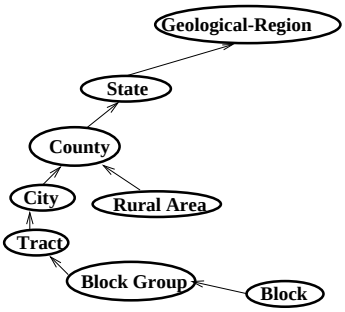
An ontology may be defined as the specification of a representational vocabulary for a shared domain of discourse which may include definitions of classes relations functions and other objects [Gruber, 1993]. A crucial concept in creating an ontology is the notion of *terminological commitment* which requires that subsribers to a given ontology agree on the semantics of any term in that ontology. This makes it incumbent upon content providers subscribing to a particular ontology to ensure that the information stored in their repositories is somehow *mapped* to the terms in the ontology. Content users on the other hand need to specify their information requests by using terms from the same ontology. A terminological commitment may be achieved via various means, such as, alignment with a dominant standard or ontology, or via a negotiation process. Terminological commitments act as a bridge between various content providers and users. This is crucial as this terminological commitment then carries forward to the metadata descriptions constructed from these ontological concepts. However, in some cases, content providers and subscribers may subscribe to different ontologies, in which case terminological commitments need to be expanded to multiple

ontologies, a situation we discuss later in this chapter. We view terminological commitments as a very important requirement for capturing the semantics of domain specific terms.

For the purposes of this chapter, we assume that media types presenting related information share the same domain of discourse. Typically there may be other terms in the vocabulary which may not be dependent on the domain and may be media specific. Further it may be necessary to translate between descriptive vocabularies that involve approximating abstracting or eliminating terms as a part of the negotiated agreement reached by various content managers. It may also be important to translate domain specific terms to domain independent media specific terms by using techniques specialized to that media type. An example of a classification that can serve as a vocabulary for constructing metadata is illustrated in Figure 1.6.



A classification using a generalization hierarchy



A classification using an aggregation hierarchy

FIGURE 1.6
Hierarchies describing a Domain Vocabulary

In the process of construction, we view the ontology from the following two different dimensions:

1. *Data driven vs Application driven* dimension

   Data driven perspective This refers to the concepts and relationships designed by interactive identification of objects in the digital content corresponding to different media types.

   Application driven perspective This refers to the concepts and relationships inspired by the class of queries for which the related information in the various media types is processed The concept *Rural Area* in Figure 1.6 is one such example.

2. *Domain dependent vs Domain independent* dimension

   Domain dependent perspective This represents the concepts which are closely tied to the domain of the application we wish to model. These are likely to be identified using the application driven approach.

Domain independent perspective This represents the concepts required by the various media types e.g., color shape and texture for images, such as R features [Jain and Hampapur, 1994]) to identify the domain specific concepts These are typically independent of the application domain and are generated by using the data driven approach.

1.4.2 Controlled Vocabulary for Digital Media

In this section we survey the terminology and vocabulary identified by various researchers for characterizing multimedia content and relate the various terms to the perspectives discussed above.

Vocabulary Feature	Media Type	Domain Dependent or Independent	Application or Data Driven
Q Features (Jain and Hampapur)	Video, Image	Domain Dependent	Application Driven
R Features (Jain and Hampapur)	Video, Image	Domain Independent	Data Driven
English Words (Kiyoki et. al.)	Image	Domain Dependent	Data Driven
ISCC and NBS colors (Kiyoki et. al.)	Image	Domain Independent	Data Driven
AVHRR features (Anderson and Stonebraker)	Image	Domain Independent	Data Driven
NDVI (Anderson and Stonebraker)	Image	Domain Dependent	Data Driven
Subword units (Glavitsch et. al.)	Audio, Text	Domain Dependent	Data Driven
Keywords (Chen et. al.)	Image, Audio Text	Domain Dependent	Application and Data Driven

Controlled Vocabulary for Digital Media

Jain and Hampapur [Jain and Hampapur, 1994] have used domain models to assign a qualitative label to a feature (such as *pass*, *dribble* and *dunk* in basketball) and are called Q Features. Features which rely on low level domain independent models like object trajectories are called R Features. Q Features may be considered as an example of the domain dependent application driven perspective, wherea R Features may be associated with the domain independent data driven perspective.

Kiyoki et. al. [Kiyoki et al., 1994] have used basic words from the General Basic English Dictionary as features which are then associated with the images. These features may be considered as examples of the domain dependent data driven perspective. Color names defined by ISCC (Inter Society Color Council) and NBS (National Bureau of Standard) are used as features, and may be considered as examples of the domain independent data driven perspective.

Anderson and Stonebraker [Anderson and Stonebraker, 1994] model some features that are primarily based on the measurements of channels Advanced Very High Resolution Radiometer (AVHRR) channels. Other features refer to spatial latitude longitude and temporal (begin date, end date) information. These may be considered as examples of domain independent data driven perspective. However there are features like the normalized difference vegetation index (NDVI) which are derived from different channels, and may be considered as an example of the domain dependent data driven perspective.

Glavitsch et. al. [Glavitsch et al., 1994] have determined from experiments that good indexing features lay between phonemes and words. They have selected three special types of subword

units VCV-, CV- and VC-. The letter V stands for a maximum sequence of vowels and C for a maximum sequence of consonants. They process a set of speech and text documents to determine a vocabulary for the domain. The same vocabulary is used for both speech and text media types, and may be considered as examples of the domain dependent data driven perspective.

Chen et. al. [Chen et al., 1994] use the keywords identified in text and speech documents as their vocabulary. Issues of restricted vs unrestricted vocabulary are very important. These may be considered as examples of the domain dependent data and application driven perspectives A summary of the above discussion is presented in Table 1.4.2.

1.4.3 Ontology guided Metadata Extraction

The extraction of metadata from the information in various media types can be primarily guided by the domain specific ontology though it may also involve terms in the domain independent ontology.

Kiyoki et. al. [Kiyoki et al., 1994] describe the automatic extraction of impression vectors based on English words or ISCC and NBS colors. The users when querying an image database then use English words to query the system One way of guiding the users could be to display the list of English words used to construct the metadata in the first place. Glavitsch et. al. [Glavitsch et al., 1994] describe the construction of a speech feature index for both text and audio documents based on a common vocabulary consisting of subword units. Chen et. al. [Chen et al., 1994] describe the construction of keyword indices, topic change indices and layout indices. These typically depend on the content of the documents and the vocabulary is dependent on keywords present in the documents.

In the above cases the vocabulary is not pre defined and depends on the content of the documents in the collection. Also the interrelationships between the terms in the ontology is not identified. A controlled vocabulary with terms and their interrelationships can be exploited to create metadata which model domain dependent relationships as illustrated by the GIS example discussed in [Kashyap and Sheth, 1997].

Example: Consider a decision support query across multiple data repositories possibly representing data in multiple media.

Get all regions having a population greater then 500, area greater than 50 acres having an urban land-cover and moderate relief.

The metadata (referred to as m-context) can be represented as:

(**AND** region (population $>$ 500) (area $>$ 50) (= land-cover “urban”) (= relief “moderate”))

Suppose the ontology from which the metadata description is constructed supports complex relationships. Furthermore, let:

CrowdedRegion $\equiv$ (**AND** region (population $>$ 200))

Inferences supported by the ontology enable determination that the regions required by the query metadata discussed earlier are instances of CrowdedRegion. Thus the metadata description (now referred to as c-context) can be rewritten as:

(**AND** CrowdedRegion (population $>$ 500) (area $>$ 50) (= land-cover “urban”) (= relief “moderate”))

The above example illustrates how metadata expressions, when constructed using ontological concepts, can take advantage of ontological inferences to support metadata computation.

1.4.4 Medical Vocabularies and Terminologies: The UMLS Project

Metadata descriptions constructed from controlled vocabularies have been used extensively to index and search for information in medical research literature. In particular, the articles in the PubMed bibliographic database has used terms obtained from the MeSH vocabulary to annotate medical research articles. Besides this there are a wide variety of controlled vocabularies in medicine used capture information related to disesases, drugs, laboratory tests, etc. Efforts have been made to integrate various perspectives by creating a “Meta” Thesaurus or vocabulary that links these vocabularies together. This was the goal of the Unified Medical Language System (UMLS) project,

initiated in 1986 by the U.S. National Library of Medicine (NLM) [Lindbergh et al., 1993]. The UMLS consists of biomedical concepts and associated strings (Metathesaurus), a semantic network and a collection of lexical tools and has been used in a large variety of applications. The three main Knowledge Sources in the UMLS are:

1. The UMLS Metathesaurus provides a common structure for more than 95 source biomedical vocabularies, organized by concept or meaning. A concept is defined as a cluster of terms (one or more words representing a distinct concept) representing the same meaning (e.g., synonyms, lexical variants, translations). The 2002 version of the Metathesaurus contains 871,584 concepts named by 2.1 million terms. Inter concept relationships across multiple vocabularies, concept categorization, and information on concept co-occurrence in MEDLINE are also included [McCray and Nelson, 1995].

2. The UMLS Semantic Network categorizes Metathesaurus concepts through semantic types and relationships [McCray and Nelson, 1995]

3. The SPECIALIST lexicon contains over 30,000 English words, including many biomedical terms. Information for each entry, including base form, spelling variants, syntactic category, inflectional variation of nouns and conjugation of verbs, is used by the lexical tools [McCray et al., 1994]. There are over 163,000 records in the 2002 SPECIALIST lexicon representing over 268,000 distinct strings.

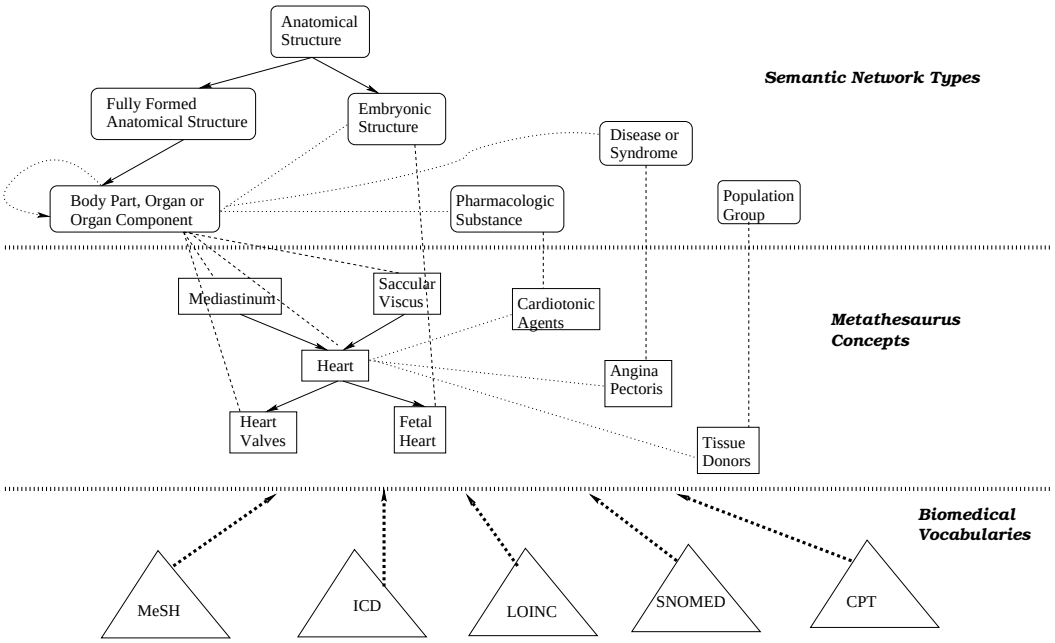


FIGURE 1.7
Biomedical vocabularies and the Unified Medical Language System

Some of the prominent medical vocabularies are as follows:

Medical Subject Headings (MeSH) The Medical Subject Headings (MeSH) [Nelson et al., 2001] have been produced by the National Library of Medicine (NLM) since 1960. The MeSH thesaurus is NLM's controlled vocabulary for subject indexing and searching of journal articles

in PubMed, and books, journal titles, and non-print materials in NLM's catalog. Translated into many different languages, MeSH is widely used in indexing and cataloging by libraries and other institutions around the world. An example of the MeSH expression used to index and search for the concept "Mumps pancreatitis" is illustrated in Figure 1.8

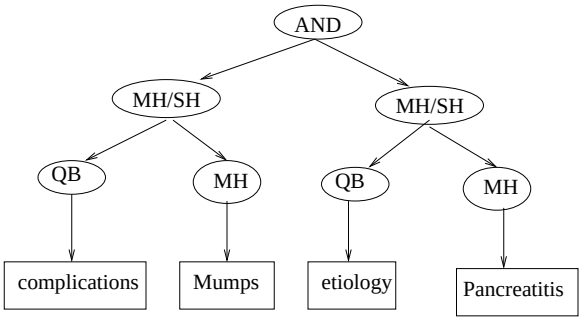


FIGURE 1.8
A MeSH descriptor for Information Retrieval

International Classification of Diseases (ICD) The International Classification of Diseases, 9th Revision (ICD-9)[ICD] is designed for the classification of morbidity and mortality information for statistical purposes, for the indexing of hospital records by disease and operations, and for data storage and retrieval. ICD-9-CM is a clinical modification of the World Health Organization's International Classification of Diseases, 9th Revision (ICD-9). The term "clinical" is used to emphasize the modification's intent: to serve as a useful tool in the area of classification of morbidity data for indexing of medical records, medical care review, and ambulatory and other medical care programs, as well as for basic health statistics. To describe the clinical picture of the patient, the codes must be more precise than those needed only for statistical groupings and trend analysis.

Systematized Nomenclature for Medicine (SNOMED) The SNOMED [Snomed] vocabulary was designed to address the need for a detailed and specific nomenclature to accurately reflect, in computer readable format, the complexity and diversity of information found in a patient record. The design ensures clarity of meaning, consistency in aggregation and ease of messaging. The SNOMED is compositional in nature, i.e., new concepts can be created as compositions of existing ones, and has a systematized hierarchical structure. It's unique design allows for the full integration of electronic medical record information into a single data structure. Overall, SNOMED has contributed to the improvement in patient care, reduction of errors inherent in data coding, facilitation of research and support of compatibility across software applications.

Current Procedural Terminology (CPT) The Current Procedural Terminology (CPT) codes [CPT] are used to describe services in electronic transactions. CPT was developed by the American Medical Association (AMA) in the 1960s, and soon became part of the standard code set for Medicare and Medicaid. In subsequent decades, it was also adopted by private insurance carriers and managed care companies, and has now become the de facto standard for reporting health care services.

Logical Observation Identifier Names and Codes (LOINC) The purpose of the Logical Observation Identifier Names and Codes (LOINC) database [LOINC] is to facilitate the exchange

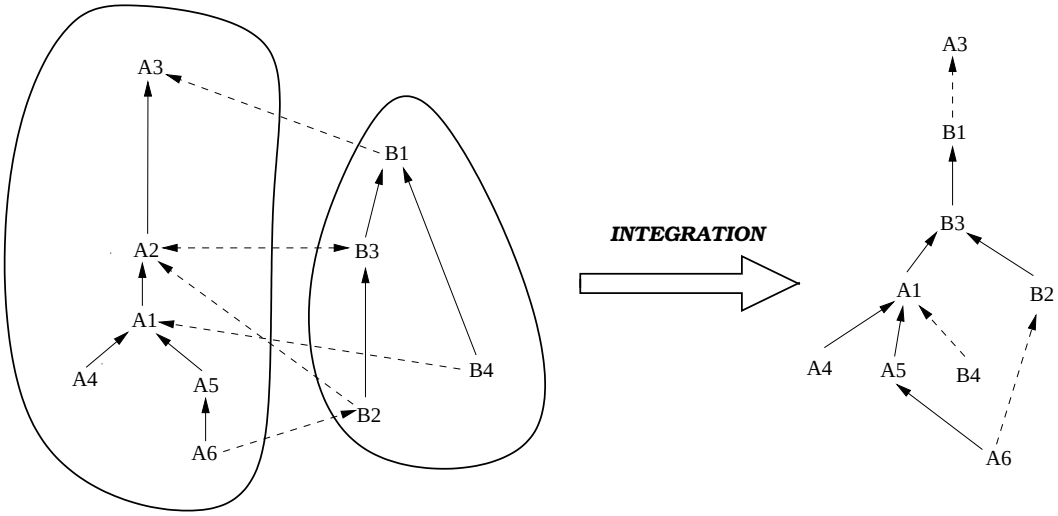


FIGURE 1.9
Expanding Terminological Commitments by Integration of Ontologies

and pooling of results, such as blood hemoglobin, serum potassium, or vital signs, for clinical care, outcomes management, and research. Its purpose is to identify observations in electronic messages such as Health Level Seven (HL7) [HL7] observation messages, so that when hospitals, health maintenance organizations, pharmaceutical manufacturers, researchers, and public health departments receive such messages from multiple sources, they can automatically file the results in the right slots of their medical records, research, and/or public health systems.

1.4.5 Expanding Terminological Commitments across multiple Ontologies

We discussed in the beginning of this section, the desirability of expanding the process of achieving terminological commitments across multiple ontologies. The UMLS system described above may be viewed as an attempt to establish terminological commitments against a multitude of biomedical vocabularies. The UMLS Metathesaurus may be viewed as a repository of inter-vocabulary relationships. Establishing terminological commitments across users of the various biomedical vocabularies would require using the relationships represented in the UMLS Metathesaurus to provide translations from a term in a source vocabulary to a term or expression of terms in a target vocabulary. This requires the ability to integrate the two vocabularies in a common graph structure and navigation of the graph structure for suitable translation. This is illusrated in an abstract manner in Figure 1.9 and is being investigate in the context of the Semantic Vocabulary Interoperation Project at the NLM [SVIP].

1.5 Conclusions

The success of the World Wide Web has led to the availability of tremendous amounts of heterogeneous digital content. However, this has led to concerns to the scalability and information loss (e.g.,

loss in precision/recall). Information modeling is viewed as an approach for enabling the scalable development of the web which would enable access to information in an information preserving manner. Creation and extraction of machine understandable metadata is a critical component of the Semantic Web effort which aims at enhancing the current web with the "semantics" of the information.

In this chapter we presented a discussion on metadata, it's use in various applications having relevance to the web and a classification of various metadata types capturing different levels of information content. We discussed approaches that use metadata descriptions for creating information models and spaces and various ways by which the attempt to capture the semantics of the information embedded in the data. In this context, we also discussed the role played by controlled vocabularies and ontologies in providing the reference terms and concepts for constructing metadata descriptions. Examples from the domain of biomedical information were presented and issues related to the establishment of terminological commitments across multiple user communities were also discussed. The role played by metadata and ontologies is crucial in modeling information and semantics, and this chapter provides an introduction to these technologies from that perspective.